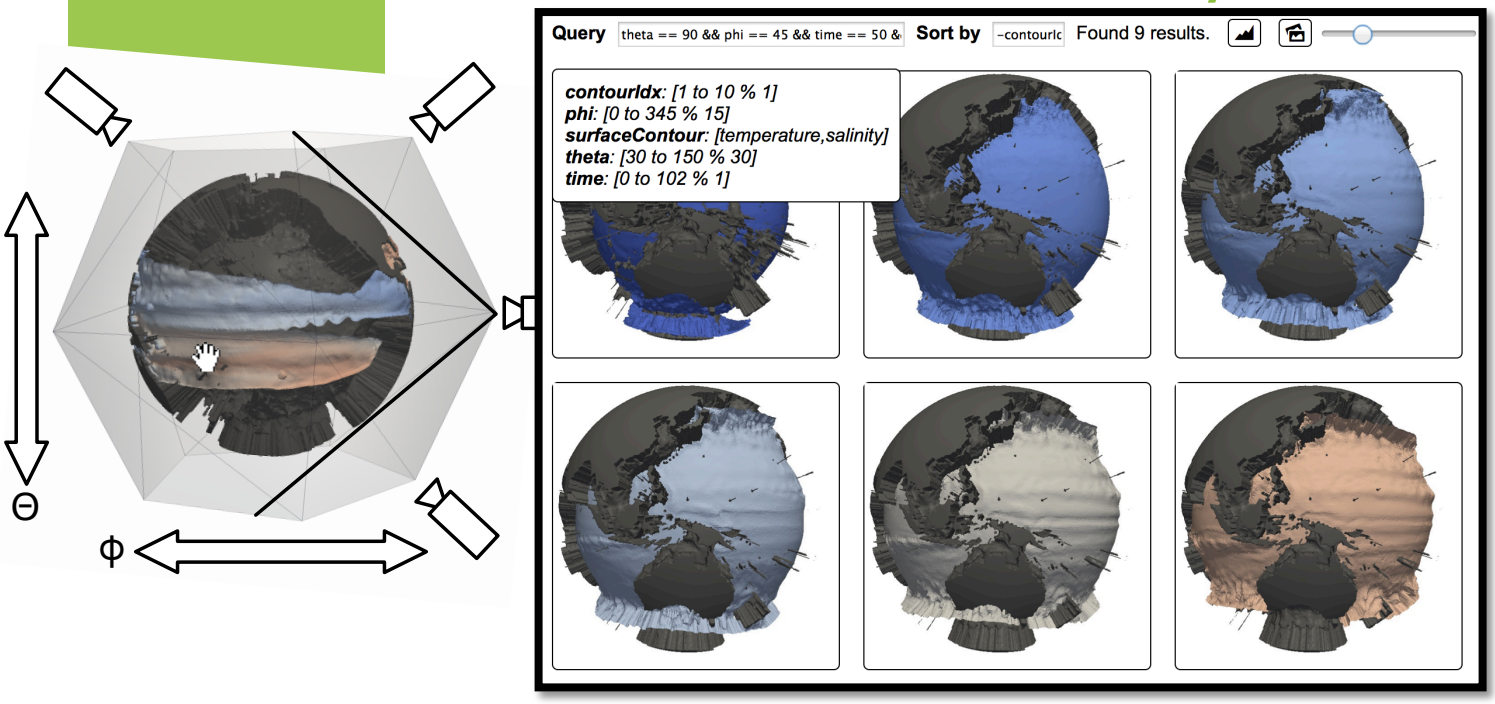
# Exploration of Exascale *In Situ* Visualization and Analysis




2015

New visualization and analysis approaches are required at exascale. We prototype and analyze two ends of *in situ* spectrum in order to evaluate them for their effectiveness.


PI: James Ahrens
Co-I: Jon Woodring
Collaborators: Joanne Wendelberger, Kary Myers, David Rogers, John Patchett, Francesca Samsel, Ayan Biswas, Yu Su, Boonthanome Nouanesengsy



**Abstract**

DOE Office of Science has embarked upon a revolutionary path to exascale ($10^{18}$ FLOPS) supercomputing. Just as the massive computing power of these machines will improve how we do simulation science, these new machines will change how we analyze simulation results. Typically at smaller scales, the data are stored or moved to another machine for post-processing visual analysis. The fundamental exascale data analysis challenge is that there is too much simulation data and too little transfer bandwidth.

*In situ* techniques that process the data while it still resides in simulation memory show a promising path forward. We see the goals of *in situ* visualization and analysis operations are multifaceted: 1) to identify important elements of the simulations, 2) to significantly reduce the data needed to preserve these elements, and 3) to offer as much flexibility as possible for future post- processing exploration.

To achieve these goals, we propose to explore two differing *in situ* approaches. The first approach, reduced simulation data, down-samples the full resolution simulation data, producing subsampled data for later post-processing visualization. The second approach, an image database, down-samples the set of input visualization and analysis parameters to produce a limited set of imagery. We are interested in exploring these two *in situ* approaches, because they are at extreme ends of a spectrum. It will allow us to understand the advantages and disadvantages of *in situ* analysis and we will evaluate both of the approaches on their effectiveness. Using this evaluation, we will merge the best of both approaches to produce an optimized *in situ* exascale visualization and analysis approach.



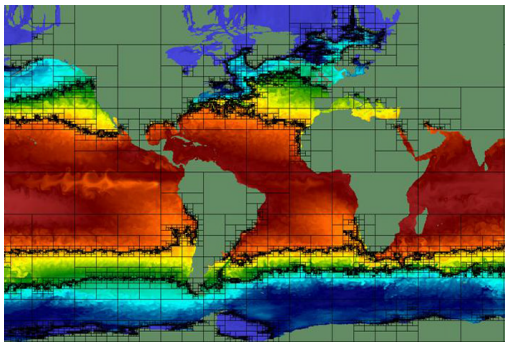
LOS ALAMOS NATIONAL LAB
Computer, Computational and Statistical Sciences Division
Los Alamos, NM 87545

PHONE
505-667-5797

WEB
http://datascience.lanl.gov

## Accomplishments and Impact on Exascale:

## Reduced Simulation Data Approach



***Advanced Sampling Strategies -*** To be able to cope with exascale data volumes, our approach is to treat it as an explicit sampling problem. We must purposefully sample the output data of large-scale simulations and experiments in time, space, and variable. We require a notion of what is "the important data" so that we can keep the most relevant data under constraints. Analysis-Driven Refinement (ADR) is our *in situ* framework for prioritizing large-scale data sets, with the goal of automatically deciding which data to keep and which to throw away. It does so by partitioning data across multiple dimensions based on user-specified importance criteria while the data set is being generated. Example prioritization algorithms include statistical and information theoretical measurements, while others include domain-specific ranking. Then, analysis specifies what actions to take with prioritized data, such as automatic camera placement, data sampling and adaptive compression.

B. Nouanesengsy, J. Woodring, K. Myers, J. Patchett, and J. Ahrens "ADR Visualization: A Generalized Framework for Ranking Large-Scale Scientific Data using Analysis-Driven Refinement", LDAV 2014, November 2014, Paris, France.

K. Myers, E. Lawrence, M. Fugate, J. Woodring, J. Wendelberger, and J. Ahrens. "An *In Situ* Approach for Approximating Complex Computer Simulations and Identifying Important Time Steps", in submission, arXiv:1409.0909 [stat.ME].

Biswas, S. Dutta, H.-W. Shen, J. Woodring. "An Information-Aware Framework for Exploring Multivariate Data Sets." IEEE Visualization 2013, Atlanta, GA, November, 2013.

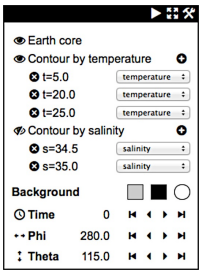
***Improved Sampling Strategies through Studying Error Effects -*** Bitmap indexing is a powerful method to organize large-scale data sets for sub-setting via queries. Our research has found that it can be leveraged as a way to sample large-scale data sets in an introspective way for reductions. This is because a bitmap index serves as a proxy for a large-scale data set that describes the properties of it, in an efficient and compact way. For example, statistical information can be extracted out of just a bitmap index, rather than the full data set, such as distributions and correlations. Therefore, we can utilize it for advanced data reduction strategies such as data sampling with quantified error bounds.

B Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger and J. Ahrens, "Effective and Efficient Data Sampling Using Bitmap Indices", Cluster Computing, March 2014.

Y. Su, G. Agrawal, J. Woodring, A. Biswas and H.-W. Shen, "Supporting Correlation Analysis on Scientific Datasets in Parallel and Distributed Settings", in Proceedings of the International ACM Symposium on High-Performance Parallel and Distribued Computing (HPDC'14), June 2014, Vancouver, Canada.

Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger and J. Ahrens. "Taming Massive Distributed Datasets: Data Sampling Using Bitmap Indices." In Proceedings of the International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'13), New York, NY, USA, June 2013.

Y. Su, G. Agrawal, and J. Woodring, "Indexing and Parallel Query Processing Support for Visualizing Climate Datasets", Proceedings of the 41st International Conference on Parallel Processing, Pittsburgh, PA, Sept. 2012.



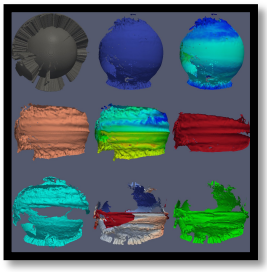



**Image Database Approach -** Imagery is on the order of 10^6 in size, whereas extreme scale simulation data is on the order of 10^15 or 10^18 in size. Our image-based approach reduces the simulation output by storing a set of output images directly from the simulation into an image database. One can think of this approach as the traditional *in situ* mode, but we are sampling the visualization and analysis parameter space, such as camera positions, operations, parameters to operations, etc., to produce a set of images stored in a database. It's important to note these images are derived from full-resolution data with high accuracy. For example, suppose we have an extreme scale simulation that calculates temperature and density over 1000 of time steps. For both variables, a scientist would like to visualize 10 isosurface values and X, Y, and Z cut planes for 10 locations in each dimension. One hundred different camera positions are also selected, in a hemisphere above the dataset pointing towards the data set. We will run the *in situ* image acquisition for every time step. These parameters will produce: 2 variables X 1000 time steps X (10 isosurface values + 3 X 10 cut planes) X 100 camera positions X 3 images (depth, float, and lighting) = 2.4 x 10 ^ 7 images. If we assume each image is 1MB (megapixel, four byte image), this results in approximately 24 TBs, which is a reasonable size for a large simulation.

J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, M. Petersen, "An Image-based Approach to Extreme Scale *In Situ* Visualization and Analysis", Supercomputing 2014, New Orleans.